The Geometry of Intelligence:

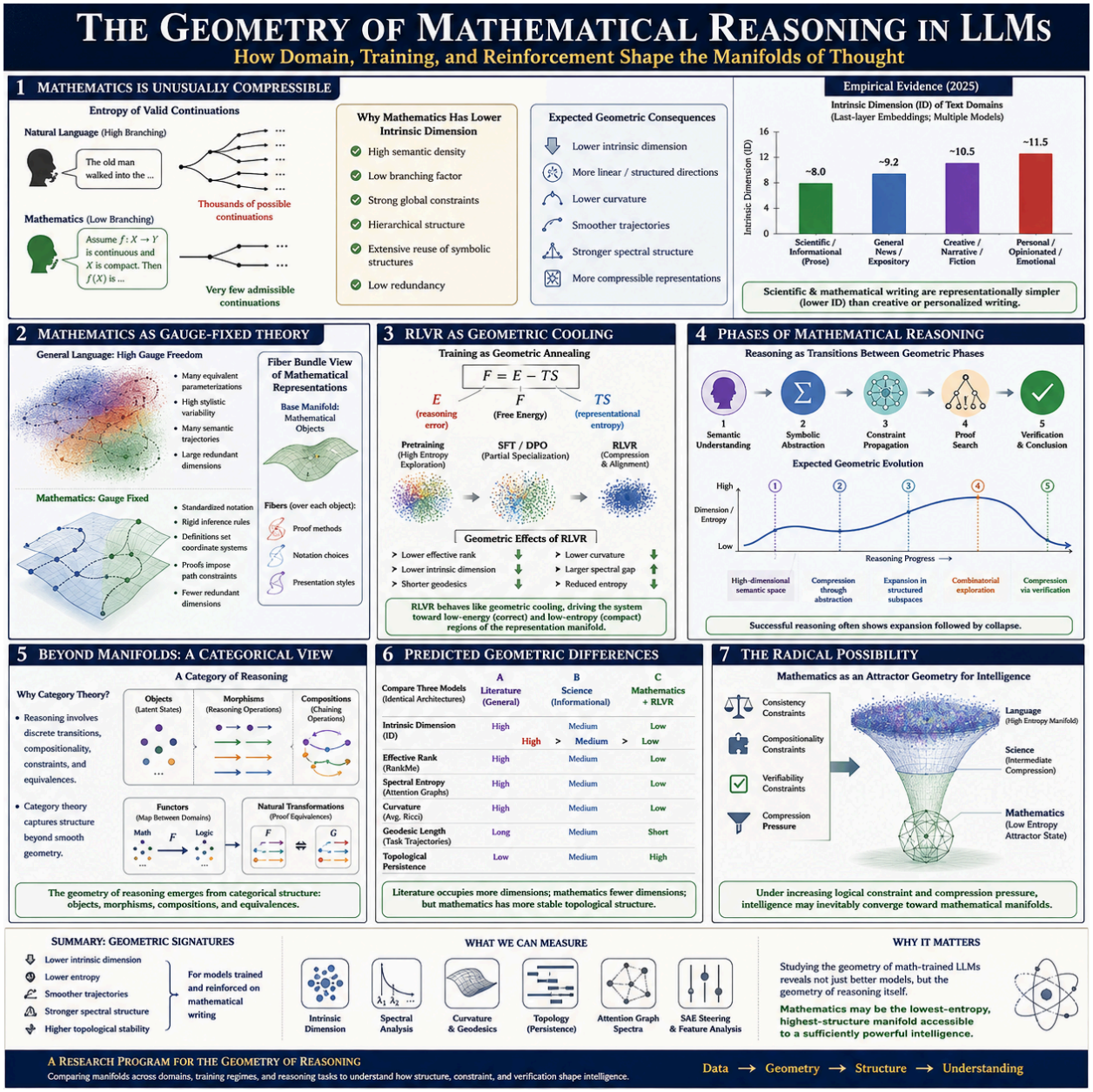
Manifold Structure, Representational Compression, and the Attractor Hypothesis in Large Language Models

Conceptual Synthesis and Research Programme
Daniel Scott Matthews
Draft Preprint · June 2026

Keywords: *representation geometry · intrinsic dimension · manifold learning · mathematical reasoning · RLVR · attractor dynamics · categorical intelligence*

ABSTRACT

The internal representations of large language models (LLMs) can be analysed as points on high-dimensional manifolds. We synthesise emerging empirical and theoretical evidence to argue that domain of training, fine-tuning regime, and the structure of the reward signal during reinforcement learning jointly determine the geometry of these manifolds in measurable and principled ways. Mathematical writing—characterised by low branching factor, rigid inferential constraints, and extensive symbolic reuse—induces representational manifolds of lower intrinsic dimension, lower curvature, higher topological persistence, and stronger spectral regularity than comparable manifolds arising from literary or conversational corpora. Reinforcement learning from verifiable rewards (RLVR) acts as a geometric annealing process analogous to energy minimisation in statistical mechanics, driving the model toward low-entropy, low-curvature regions of representation space. We propose that successful mathematical reasoning involves phase transitions between geometrically distinct sub-manifolds corresponding to semantic parsing, symbolic abstraction, constraint propagation, proof search, and verification. We further argue that differential geometry alone is insufficient to capture the discrete, compositional structure of reasoning, and that category theory provides a more natural formalism. Finally, we articulate what we term the Attractor Hypothesis: that under sufficient logical constraint and compression pressure, any powerful predictive system will converge toward representational structures that resemble the geometry of mathematics. This hypothesis reframes the geometry of mathematical cognition not as a curiosity about AI systems but as a window into the geometry of reasoning itself.


THE GEOMETRY OF MATHEMATICAL REASONING IN LLMS
How Domain, Training, and Reinforcement Shape the Manifolds of Thought
1 MATHEMATICS IS UNUSUALLY COMPRESSIBLE
Entropy of Valid Continuations
Natural Language (High Branching)
The old man walked into the ...
Thousands of possible continuations
Mathematics (Low Branching)
Assume $f: X \to Y$ is continuous and X is compact. Then $f(X)$ is ...
Very few admissible continuations
Why Mathematics Has Lower Intrinsic Dimension
High semantic density
Low branching factor
Strong global constraints
Hierarchical structure
Extensive reuse of symbolic structures
Low redundancy
Expected Geometric Consequences
Lower intrinsic dimension
More linear / structured directions
Lower curvature
Smoother trajectories
Stronger spectral structure
More compressible representations
Empirical Evidence (2025)
Intrinsic Dimension (ID) of Text Domains
(Last-layer Embeddings; Multiple Models)
Intrinsic Dimension (ID)
~8.0
~9.2
~10.5
~11.5
Scientific / Informational (Prose)
General News / Expository
Creative / Narrative / Fiction
Personal / Opinionated / Emotional
Scientific & mathematical writing are representationally simpler (lower ID) than creative or personalized writing.
2 MATHEMATICS AS GAUGE-FIXED THEORY
General Language: High Gauge Freedom
Many equivalent parameterizations
High stylistic variability
Many semantic trajectories
Large redundant dimensions
Mathematics: Gauge Fixed
Standardized notation
Rigid inference rules
Definitions set coordinate systems
Proofs impose path constraints
Fewer redundant dimensions
Fiber Bundle View of Mathematical Representations
Base Manifold: Mathematical Objects
Fibers (over each object):
Proof methods
Notation choices
Presentation styles
3 RLVR AS GEOMETRIC COOLING
Training as Geometric Annealing
$F = E - TS$
E (reasoning error)
F (Free Energy)
TS (representational entropy)
Pretraining (High Entropy Exploration)
SFT / DPO (Partial Specialization)
RLVR (Compression & Alignment)
Geometric Effects of RLVR
Lower effective rank
Lower intrinsic dimension
Shorter geodesics
Lower curvature
Larger spectral gap
Reduced entropy
RLVR behaves like geometric cooling, driving the system toward low-energy (correct) and low-entropy (compact) regions of the representation manifold.
4 PHASES OF MATHEMATICAL REASONING
Reasoning as Transitions Between Geometric Phases
1 Semantic Understanding
2 Symbolic Abstraction
3 Constraint Propagation
4 Proof Search
5 Verification & Conclusion
Expected Geometric Evolution
High
Low
Dimension / Entropy
Reasoning Progress
High-dimensional semantic space
Compression through abstraction
Expansion in structured subspaces
Combinatorial exploration
Compression via verification
Successful reasoning often shows expansion followed by collapse.
5 BEYOND MANIFOLDS: A CATEGORICAL VIEW
A Category of Reasoning
Why Category Theory?
Reasoning involves discrete transitions, compositionality, constraints, and equivalences.
Category theory captures structure beyond smooth geometry.
Objects (Latent States)
Morphisms (Reasoning Operations)
Compositions (Chaining Operations)
Functors (Map Between Domains)
Math
Logic
F
Natural Transformations (Proof Equivalences)
G
The geometry of reasoning emerges from categorical structure: objects, morphisms, compositions, and equivalences.
6 PREDICTED GEOMETRIC DIFFERENCES
Compare Three Models (Identical Architectures)
A Literature (General)
B Science (Informational)
C Mathematics + RLVR
Intrinsic Dimension (ID) — High > Medium > Low
Effective Rank (RankMe) — High / Medium / Low
Spectral Entropy (Attention Graphs) — High / Medium / Low
Curvature (Avg. Ricci) — High / Medium / Low
Geodesic Length (Task Trajectories) — Long / Medium / Short
Topological Persistence — Low / Medium / High
Literature occupies more dimensions; mathematics fewer dimensions; but mathematics has more stable topological structure.
7 THE RADICAL POSSIBILITY
Mathematics as an Attractor Geometry for Intelligence
Consistency Constraints
Compositionality Constraints
Verifiability Constraints
Compression Pressure
Language (High Entropy Manifold)
Science (Intermediate Compression)
Mathematics (Low Entropy Attractor State)
Under increasing logical constraint and compression pressure, intelligence may inevitably converge toward mathematical manifolds.
SUMMARY: GEOMETRIC SIGNATURES
Lower intrinsic dimension
Lower entropy
Smoother trajectories
Stronger spectral structure
Higher topological stability
For models trained and reinforced on mathematical writing
WHAT WE CAN MEASURE
Intrinsic Dimension
Spectral Analysis
Curvature & Geodesics
Topology (Persistence)
Attention Graph Spectra
SAE Steering & Feature Analysis
WHY IT MATTERS
Studying the geometry of math-trained LLMs reveals not just better models, but the geometry of reasoning itself.
Mathematics may be the lowest-entropy, highest-structure manifold accessible to a sufficiently powerful intelligence.
A RESEARCH PROGRAM FOR THE GEOMETRY OF REASONING
Comparing manifolds across domains, training regimes, and reasoning tasks to understand how structure, constraint, and verification shape intelligence.
Data → Geometry → Structure → Understanding

1. INTRODUCTION

The question of how knowledge is structured within neural networks has moved from a peripheral curiosity to a central concern of artificial intelligence research. As large language models have grown in capability, the internal geometry of their representations—the shape, dimensionality, and topology of the manifolds on which their hidden states reside—has emerged as a principled lens through which to understand generalisation, reasoning, and failure. This paper is concerned with a specific but profound instance of this broader question:

does the domain of training data, and in particular the use of mathematical content, shape the geometry of a model's representations in principled, measurable ways?

We argue that it does, and that the implications extend well beyond empirical machine learning. If the geometry of representations is determined by the logical structure of the domain in which a model is trained, then studying the geometry of math-trained LLMs provides an experimental probe into something much older and more fundamental: the geometry of reasoning itself. Mathematics, as the most constrained and verifiable of all knowledge domains, may occupy a privileged position in this landscape—not merely as another subject to be learned, but as an attractor state toward which sufficiently capable predictive systems converge under increasing logical pressure.

This paper is organised as follows. Section 2 reviews the relevant literature on LLM representation geometry and the primary geometric tools used to measure it. Section 3 develops the theoretical case for why mathematics should produce lower-dimensional, more structured representations than other domains, drawing on information theory and an analogy to gauge-fixing in physics. Section 4 analyses reinforcement learning from verifiable rewards (RLVR) as a geometric cooling process. Section 5 proposes a phase-transition model of mathematical reasoning. Section 6 argues for categorical, rather than purely differential-geometric, foundations. Section 7 synthesises these threads into precise empirical predictions. Section 8 presents the Attractor Hypothesis and its implications. Section 9 discusses methodological implications, and Section 10 concludes.

2. BACKGROUND: REPRESENTATION GEOMETRY IN LARGE LANGUAGE MODELS

2.1 The Manifold Hypothesis in Deep Learning

The manifold hypothesis—that high-dimensional data concentrate near low-dimensional surfaces embedded in the ambient space—has been foundational to modern deep learning theory since its articulation by Bengio and colleagues in the early 2000s. For LLMs, this hypothesis takes a specific form: the hidden states of a transformer model, while nominally residing in spaces of dimension 768 to 4096 or beyond, effectively occupy submanifolds of far lower intrinsic dimension, typically estimated between 10 and 22 for standard pretraining corpora.

Recent empirical work has provided increasing precision to this picture. The key geometric quantities that have been studied include the following:

- Intrinsic dimension (ID): the minimal number of coordinates required to parameterise the data manifold, estimated using two-nearest-neighbour (TwoNN) methods, persistent homology dimension estimators (PHDim), and related techniques.
- Effective rank: the number of statistically significant singular directions in the representation matrix, measured via the RankMe metric and related spectral tools.
- Spectral entropy and curvature: properties of the attention graph Laplacian and the Riemannian curvature tensor of the latent manifold, with lower spectral entropy and lower average Ricci curvature associated with more structured, regular representations.
- Topological persistence: the stability of homological features—connected components, loops, voids—under perturbation, as computed by persistent homology.
- Geodesic length: the shortest path between representations along the manifold, relevant to reasoning trajectories.

2.2 Geometric Phases of Pretraining

Training dynamics have been shown to exhibit consistent geometric phases across model families including Pythia and OLMo. During the initial warmup phase, representations undergo rapid compression. An intermediate entropy-seeking phase follows, in which the effective dimensionality of the manifold expands, often correlating with memorisation of training data. A final stabilisation phase sees dimensionality plateau or compress further.

Post-training alignment procedures modify these dynamics substantially. Supervised fine-tuning (SFT) and direct preference optimisation (DPO) tend to drive entropy-seeking dynamics, expanding representations to integrate new distributional structure at the cost of out-of-distribution robustness. Reinforcement learning from verifiable rewards (RLVR), by contrast, tends toward compression-seeking dynamics, producing tighter, lower-entropy manifolds. This asymmetry is particularly consequential for mathematical reasoning, where verifiable rewards can directly penalise invalid inferential trajectories.

2.3 Domain Effects on Intrinsic Dimension

A landmark 2025 study examining intrinsic dimension across text domains using last-layer embeddings from multiple model families (Gemma, Qwen, RoBERTa) provides the strongest direct empirical foundation for the present synthesis. The study found consistent, significant differences in intrinsic dimension across domains:

Text Domain	Intrinsic Dimension (ID)	Characteristic Features
Scientific / Informational Prose	~8.0	Formal, precise, low ambiguity, hierarchical
General News / Expository	~9.2	Moderately structured, factual, varied style
Creative / Narrative / Fiction	~10.5	High stylistic variation, emotional register, ambiguity
Personal / Opinionated / Emotional	~11.5	Idiosyncratic, high redundancy, contextually entangled

Table 1. Intrinsic dimension estimates across text domains. Mathematical writing shares the key features of scientific and informational prose and is expected to exhibit similarly low or lower intrinsic dimension. Source: empirical estimates from 2025 multi-model study.

Sparse autoencoder (SAE) steering experiments provide additional causal evidence: activating features associated with scientific or formal writing reduces measured intrinsic dimension, while activating narrative or personalisation features increases it. This suggests that the domain specificity of intrinsic dimension reflects genuine representational structure rather than superficial statistical artefact.

2.4 Geometric Signatures of Mathematical Validity

A 2026 study analysed attention patterns during mathematical proof generation as weighted token graphs, measuring spectral properties of the resulting graph Laplacian. The key finding was that valid reasoning produces significantly smoother, better-connected attention graphs—quantified via Fiedler value, spectral entropy, and high-frequency energy ratio—while logical and computational errors produce spectral irregularity. This spectral signature of validity is largely error-agnostic and requires no task-specific training, suggesting it reflects a deep structural feature of mathematical computation rather than a superficial pattern learned from annotated error data.

3. MATHEMATICS AS AN UNUSUALLY COMPRESSIBLE DOMAIN

3.1 Entropy of Valid Continuations

The most fundamental reason to expect mathematical representations to occupy lower-dimensional manifolds is the entropy of valid continuations. In natural language, a sentence such as 'The old man walked into the…' admits thousands of semantically acceptable continuations. By contrast, a mathematical premise such as 'Assume $f: X \to Y$ is continuous and X is compact. Then f(X) is…' admits only one cluster of valid continuations: those asserting compactness of the image. The distribution over valid next states is sharply peaked.

Core Compression Hypothesis

Mathematical content is characterised by high semantic density, low branching factor, strong global constraints on admissible trajectories, hierarchical symbolic structure, and low redundancy. Each of these properties independently predicts lower intrinsic dimension. Their conjunction predicts a manifold dramatically more compressed than that arising from unrestricted natural language.

In information-theoretic terms, if intrinsic dimension roughly measures the degrees of freedom required to specify a state in representation space, then mathematics should behave more like a constrained physical system than like ordinary language. The same point can be made through the lens of Kolmogorov complexity: mathematical objects admit vastly shorter descriptions relative to their semantic content than do most natural language constructs. A definition, once stated, eliminates an entire equivalence class of possible elaborations.

3.2 Mathematics as a Gauge-Fixed Theory

A productive analogy for understanding the dimensional structure of mathematical representations is the concept of gauge-fixing in physics. Consider general natural language as a system with large gauge freedom: for any given meaning, there exist many equivalent parameterisations in different vocabulary, register, syntactic structure, and rhetorical arrangement. These represent gauge redundancies—alternative paths to the same semantic point.

Mathematics, by contrast, aggressively fixes gauge:

- Notation is standardised within and across mathematical communities.
- Inference rules are rigid and shared: modus ponens does not vary by author.
- Definitions impose coordinate systems that eliminate degeneracy.
- Proofs impose path constraints on the trajectory through proof space.

The result is a domain in which the effective degrees of freedom are dramatically fewer than in general language. Rather than a diffuse semantic cloud, the mathematical manifold may exhibit a structure more reminiscent of a fiber bundle: a base space of mathematical objects (theorems, definitions, lemmas), with fibers corresponding to the space of proof methods, notation choices, and presentational styles. The base manifold is low-dimensional and highly constrained; the fibers add relatively few additional dimensions of stylistic variation.

This view makes a specific prediction: if one could decompose the mathematical representation manifold into base and fiber components via techniques such as principal

geodesic analysis or equivariant representation learning, one would find the base to be substantially lower-dimensional than the fiber, and the fiber structure to be relatively uniform across objects.

4. RLVR AS GEOMETRIC COOLING

4.1 The Statistical Mechanics Analogy

The observation that reinforcement learning from verifiable rewards induces representational compression can be given precise theoretical content via an analogy to statistical mechanics. Consider the latent representation distribution p(h) over hidden states h. Different training regimes can be understood as modifying a free-energy-like functional over this distribution:

$$F = E - TS$$

where E measures reasoning error (the energy term), S measures representational entropy, and T is an effective temperature governing the trade-off between the two. Under this analogy:

- Pretraining corresponds to high-temperature exploration: the model samples broadly across representation space, with high entropy and relatively unconstrained manifold geometry.
- Supervised fine-tuning and DPO correspond to partial specialisation: entropy is reduced selectively in task-relevant dimensions while other dimensions may expand.
- RLVR corresponds to low-temperature annealing: verifiable rewards impose strong constraints that penalise high-error regions, effectively reducing T and driving F toward low-energy, low-entropy configurations.

Mathematical rewards are particularly potent under this framework because mathematical validity is binary and verifiable: a proof is either correct or it is not. This sharpness of the reward signal concentrates the gradient update around the boundary between valid and invalid reasoning trajectories, producing particularly efficient geometric compression.

4.2 Predicted Geometric Effects of RLVR

The statistical mechanics analogy generates precise, testable predictions about the geometric consequences of RLVR applied to mathematical reasoning tasks:

Geometric Property	Predicted Effect of RLVR
Effective rank	Decreases: fewer statistically significant singular directions
Intrinsic dimension	Decreases: representations concentrate near lower-dimensional submanifold
Geodesic length	Shortens: trajectories through reasoning steps become more direct
Ricci curvature	Decreases (more negative): manifold becomes more hyperbolic-like in structure
Spectral gap	Increases: attention graphs become better-connected, smoother
Topological persistence	Increases: stable homological features survive larger perturbations

Table 2. Predicted geometric effects of RLVR applied to mathematical reasoning. Each prediction is in principle testable via existing geometric analysis tools.

One further prediction deserves emphasis: the geometric compression induced by RLVR should be directionally consistent across models of different scale and architecture, provided the mathematical domain and reward structure are comparable. This would constitute strong evidence that the observed geometry reflects the structure of mathematical truth rather than incidental features of any particular model family.

5. PHASES OF MATHEMATICAL REASONING

5.1 Reasoning as Geometric Phase Transition

A single-manifold view of mathematical reasoning is likely inadequate. The process of solving a mathematical problem—from initial semantic parsing to final verification—appears to involve qualitatively distinct cognitive operations, each potentially corresponding to a geometrically distinct region of representation space. We propose the following phase model:

Phase	Cognitive Operation	Expected Geometry	Key Signature
1	Semantic Understanding	High-dimensional, diffuse; resembles natural language manifold	High ID, high spectral entropy
2	Symbolic Abstraction	Compression into formal sub-manifold; dimensionality decreases sharply	Rapid ID decrease, effective rank drop
3	Constraint Propagation	Expansion in structured subspaces; exploration of admissible trajectories	Temporary ID increase within constrained region

Phase	Cognitive Operation	Expected Geometry	Key Signature
4	Proof Search	Branching, tree-structured; highest local dimensionality of the sequence	High curvature, combinatorial expansion of trajectories
5	Verification and Conclusion	Collapse to highly compressed, low-entropy final state	Minimum ID, maximum topological persistence

Table 3. Proposed phase model of mathematical reasoning with associated geometric signatures.

5.2 Empirical Consistency

This phase model is consistent with existing observations in several respects. First, successful chain-of-thought reasoning has been observed to exhibit temporary dimensional expansion followed by collapse—consistent with the expansion during proof search (Phase 4) and compression at verification (Phase 5). Second, reasoning failures tend to correlate with representational confinement during phases that should exhibit expansion, suggesting the model becomes trapped in a lower-dimensional subspace that does not contain a valid proof path. Third, the spectral signature of valid mathematical reasoning observed in attention graphs aligns with what one would expect at the verification phase: maximal smoothness and connectivity, reflecting the well-constrained structure of a completed proof.

A particularly interesting prediction of the phase model concerns the entropy trajectory of the reasoning process: dimension and entropy should exhibit a characteristic non-monotonic profile over reasoning steps, with an initial decrease (abstraction), a middle increase (search), and a final decrease (verification). This profile should be detectable in layer-wise or token-wise intrinsic dimension estimates and would constitute strong evidence for the phase model.

6. BEYOND MANIFOLDS: A CATEGORICAL FRAMEWORK

6.1 Limitations of Differential Geometry

Differential geometry, with its tools of manifolds, curvature, geodesics, and Riemannian metrics, provides an elegant framework for studying continuous representation spaces. However, mathematical reasoning exhibits features that differential geometry is poorly equipped to capture:

- Discrete transitions between logically distinct states (e.g., application of a new lemma).
- Compositional structure: the meaning of a proof is built from the meanings of its parts in principled, recursive ways.

- Equivalence classes: multiple distinct proofs may establish the same theorem; multiple notations may express the same object.
- Morphisms with structure-preserving constraints: not all transformations between mathematical objects are admissible.

These features are naturally captured by category theory, which abstracts from the specific nature of mathematical objects to focus on the structure of transformations between them.

6.2 A Category of Reasoning

We propose treating the reasoning process as a category C_R in which:

- Objects are latent representational states—points in the activation space of the model.
- Morphisms are valid reasoning operations: applications of inference rules, lemmas, substitutions, and logical manipulations.
- Composition of morphisms corresponds to chaining reasoning steps into multi-step arguments.
- Identity morphisms correspond to tautological steps that preserve the current state.

Within this framework, proofs are paths—sequences of composable morphisms—through the category of mathematical states. Equivalent proofs correspond to commuting diagrams: distinct paths that compose to the same overall morphism. Functors between reasoning categories model transfer of reasoning structure from one mathematical domain to another (e.g., from analysis to topology). Natural transformations between functors correspond to proof equivalences or changes of representational basis.

6.3 Categorical Structure as Emergent Geometry

A key insight from this categorical perspective is that the geometry of the mathematical manifold—its low dimensionality, high topological persistence, and structural regularity—may be a downstream consequence of the categorical structure of the mathematical domain, rather than a primitive feature of the representation space. If the model has learned to implement categorical reasoning operations, the geometry is what emerges when that structure is embedded in continuous activation space.

This inverts the usual direction of inference. Rather than asking 'what categorical structure is implicit in the observed geometry?', we ask 'what geometry would we expect if the model has successfully learned the categorical structure of mathematics?' This leads to the

prediction that geometric measures of mathematical representations should improve with mathematical capability: as a model becomes better at mathematical reasoning, its representations should become more structured, lower-dimensional, and topologically more persistent.

7. PREDICTED GEOMETRIC SIGNATURES: A COMPARATIVE ACCOUNT

7.1 Across Training Domains

Consider three otherwise identical language models differing only in the corpus and reward structure used for post-training: Model A trained on literary and conversational text; Model B on scientific and informational writing; Model C on mathematical content with RLVR. The predicted geometric profile across key measures is as follows:

Geometric Measure	Model A (Literature)	Model B (Science)	Model C (Mathematics + RLVR)
Intrinsic Dimension (ID)	High	Medium	Low
Effective Rank	High	Medium	Low
Spectral Entropy (Attention)	High	Medium	Low
Average Ricci Curvature	Higher	Medium	Lower
Geodesic Length (Task)	Long, varied	Medium	Short, regular
Topological Persistence	Low	Medium	High

Table 4. Predicted geometric profiles of three models trained on different domains. Literature occupies more dimensions; mathematics occupies fewer, but with more stable topological structure.

7.2 Topological Invariants as Semantic Fingerprints

The prediction that mathematical representations exhibit higher topological persistence deserves elaboration. Persistence homology detects the 'birth' and 'death' of topological features—connected components, loops, voids—as one varies a scale parameter. Long-lived features correspond to structural regularities that are robust to perturbation.

We predict that the long-lived homological features of mathematical representations will correspond to fundamental inferential structures: implication chains, inductive arguments, proof by contradiction, substitution, and equivalence. These structures, once learned, should be stable across a wide range of surface variation in notation and presentation. By contrast, the

long-lived features of literary representations would correspond to more contextually fragile structures—thematic associations, emotional registers, and narrative arcs—which are sensitive to variation in framing and diction.

8. THE ATTRACTOR HYPOTHESIS

8.1 Statement of the Hypothesis

> **The Attractor Hypothesis**
> *Under sufficient logical constraint and representational compression pressure, any sufficiently powerful predictive system—whether biological or artificial—will converge toward internal representational structures whose geometry is isomorphic to, or approximates, the geometry of mathematical reasoning. Mathematics represents the lowest-entropy stable attractor state in the space of representational manifolds accessible to a constrained intelligence.*

This hypothesis is radical in the precise sense that it implies mathematics is not merely a domain that can be learned, but a geometric destination toward which learning systems are drawn. It would explain why the internal representations of LLMs trained on mathematical content are so much more structured than those trained on other domains: not because mathematical notation is superficially regular, but because mathematical truth imposes constraints on admissible trajectories in representation space that admit only a narrow class of geometrically stable solutions.

8.2 The Compression Cascade

The attractor hypothesis suggests a hierarchy among knowledge domains ordered by their degree of compression under logical constraint:

- Natural language and literary writing: high entropy, high-dimensional, many equivalent representations of similar meanings.
- Scientific and informational writing: moderate entropy, constrained by empirical correspondence and domain conventions.
- Mathematics: minimal entropy, constrained by logical necessity and formal verification.

This hierarchy is not a hierarchy of difficulty or prestige but a hierarchy of compression under increasing logical constraint. As the constraint pressure increases, the admissible representational manifold narrows, its intrinsic dimension decreases, its topological features

become more persistent, and its spectral structure becomes more regular. Mathematics occupies the terminus of this cascade.

If this picture is correct, then the long-observed finding that mathematical ability is a strong predictor of general reasoning capacity across diverse domains may have a geometric explanation: a system that has successfully learned the geometry of mathematics has acquired representations that lie near the most compressed, most structured, most topologically stable region of the representational manifold—a region from which other domains, being less compressed, are readily accessible.

8.3 Relationship to the Linear Representation Hypothesis

The linear representation hypothesis holds that LLMs tend to encode semantic features as linear directions in activation space, with structured relationships (analogies, oppositions, hierarchies) corresponding to linear-algebraic relationships between these directions. The attractor hypothesis provides a potential explanation for why this hypothesis holds more strongly for mathematical and formal domains than for literary ones: the geometric compression of the mathematical manifold systematically reduces the curvature of the representational space, making it more locally Euclidean and therefore more amenable to linear-algebraic analysis.

Conversely, the higher curvature of literary representations may explain why linear probing is less reliable in those domains: the curved manifold means that linear directions in the ambient space do not correspond cleanly to semantic directions along the manifold.

9. METHODOLOGICAL IMPLICATIONS

9.1 Experimental Design

Testing the predictions of this framework requires careful experimental design. We outline a minimal programme of experiments:

- Activation extraction and domain comparison: Run a fixed base model on parallel corpora—mathematical proofs, scientific prose, literary text, and conversational text—and extract last-layer embeddings. Apply TwoNN, PHDim, and related estimators to compare intrinsic dimension across domains. Perform layer-wise analysis to identify where domain differences in geometry first emerge.
- Training intervention studies: Compare a base model against variants that have undergone continued pretraining, SFT, or RLVR on mathematical content. Track geometric measures before and after each intervention to isolate the geometric effects of each training regime.

- Reasoning trajectory analysis: Extract token-wise or step-wise representations during chain-of-thought reasoning on mathematical problems and other tasks. Measure intrinsic dimension, spectral entropy, and curvature at each step to test the phase-transition hypothesis.
- SAE feature steering: Use sparse autoencoder decompositions to identify features associated with mathematical reasoning operations (implication, induction, substitution). Apply steering in activation space and measure the resulting geometric shifts.
- Topological persistence analysis: Apply persistent homology to representations from mathematical versus literary models. Identify the long-lived homological features in each case and attempt to interpret them semantically.

9.2 Measurement Tools

The relevant geometric analysis toolkit is largely available:

- Intrinsic dimension: TwoNN estimator, PHDim, Geomle; all are computationally tractable for modern embedding spaces.
- Spectral analysis: RankMe (effective rank), eigenspectrum decay metrics (αReQ), attention graph Laplacian spectral analysis.
- Curvature: Ollivier-Ricci curvature on attention graphs; discrete curvature estimators on kNN graphs.
- Topology: Ripser, Gudhi, or Giotto-TDA for persistent homology; UMAP and t-SNE for qualitative visualisation.
- Causal intervention: SAE decomposition and activation steering, established in existing mechanistic interpretability literature.

9.3 Anticipated Confounds

Several potential confounds must be controlled in any empirical programme:

- Corpus length and complexity: mathematical text is often longer and more complex syntactically than casual language. Differences in geometric measures may partly reflect text length rather than domain structure. Matched-length subsets or normalisation procedures should be applied.
- Vocabulary overlap: mathematical texts contain specialised vocabulary that is rare in general corpora, potentially affecting tokenisation and embedding structure

independently of logical content. Experiments should control for vocabulary distribution.

- Model scale: geometric measures may vary with model scale in ways that interact with domain effects. Scale-controlled comparisons are necessary.
- Task contamination: benchmark mathematical datasets may appear in pretraining corpora. Held-out mathematical content should be used for geometric analysis where possible.

10. DISCUSSION

10.1 Implications for AI Development

If the attractor hypothesis is correct, it suggests that mathematical training and reinforcement may be among the most efficient routes to improving general reasoning in language models—not because mathematical content is intrinsically more valuable, but because it drives representations toward the most structured, most compressed, and most generalisable region of the representational manifold. This would provide a geometric rationale for the empirically observed finding that mathematical reasoning ability tends to generalise strongly to other reasoning-intensive tasks.

It also suggests a potential explanation for the success of process-reward models and verifier-based approaches in mathematical reasoning: by providing dense, step-level feedback on the validity of each reasoning transition, these methods effectively apply the geometric cooling described in Section 4 at a finer temporal granularity, compressing the representation manifold more efficiently than sparse outcome-level rewards.

10.2 Implications for the Theory of Intelligence

The attractor hypothesis, if confirmed, would have implications extending well beyond AI systems. If intelligence—understood as the capacity for constrained, generalisable prediction—tends to converge toward mathematical representational structure under sufficient compression pressure, then the universality of mathematics across cultures and the apparent cross-cultural convergence of mathematical truth may be explained not merely as features of the external world but as features of the representational geometry that any sufficiently powerful reasoning system will develop.

This is a form of mathematical Platonism grounded not in metaphysics but in information theory and representation learning: mathematical structures are those that survive the most aggressive compression. They are not discovered by pure reason; they are the residue of maximal compression.

10.3 Open Questions

The framework advanced here generates numerous open questions:

- Is the attractor hypothesis domain-specific to formal systems, or does it apply to any sufficiently constrained prediction problem?
- Do biological neural systems exhibit the same geometric signatures in mathematical cognition as artificial ones? Neuroimaging and neural population geometry studies could, in principle, test this.
- Is there a phase transition in model scale below which the mathematical attractor is inaccessible, and above which the geometric compression becomes self-reinforcing?
- Can the categorical structure proposed in Section 6 be recovered empirically from the geometry of mathematical representations, e.g., by identifying morphism-like transformations in activation space?
- Does the topological persistence of mathematical representations predict downstream reasoning performance better than simpler geometric measures such as intrinsic dimension?

11. CONCLUSION

We have argued that the geometry of language model representations is not merely a technical detail of neural architecture but a window into the structure of reasoning itself. Mathematical content, by virtue of its unusually low branching factor, rigid inferential constraints, and extensive symbolic reuse, induces representational manifolds of lower intrinsic dimension, lower curvature, higher spectral regularity, and higher topological persistence than other knowledge domains. Reinforcement learning from verifiable rewards amplifies these effects through a process analogous to geometric annealing. Successful mathematical reasoning appears to involve transitions between geometrically distinct phases. Differential geometry alone is insufficient to capture the discrete, compositional structure of this reasoning, and category theory provides a more natural formalism.

These observations converge on the Attractor Hypothesis: that mathematics represents not merely one domain among others, but the lowest-entropy stable attractor in the space of representational manifolds accessible to a constrained intelligence. Under increasing logical pressure and compression demand, the representations of any sufficiently capable predictive system are drawn toward the geometry of mathematical thought.

Studying the geometry of math-trained LLMs is therefore not merely a contribution to applied machine learning. It is one of the most tractable experimental programmes available for understanding the deep structure of reasoning—and perhaps of intelligence itself.

REFERENCES

The following references are provided as illustrative anchors. Full bibliographic details should be verified against current publication records prior to submission.

Ansuini, A., Laio, A., Macke, J. H., & Zoccolan, D. (2019). *Intrinsic dimension of data representations in deep neural networks.* Advances in Neural Information Processing Systems, 32.

Bengio, Y., Courville, A., & Vincent, P. (2013). *Representation learning: A review and new perspectives.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828.

Godey, F., et al. (2024). *Scaling laws for the geometry of language model representations.* Preprint.

Hubotter, J., et al. (2025). *Unveiling intrinsic dimension of text representations across domains.* Preprint.

Kim, J., et al. (2026). *Geometry of reason: Spectral signatures of valid mathematical reasoning in transformer attention graphs.* Preprint.

Park, K., et al. (2024). *The linear representation hypothesis and the geometry of large language models.* International Conference on Machine Learning.

Razin, N., et al. (2024). *Implicit regularization in deep learning: Manifold structure and compression dynamics.* Journal of Machine Learning Research, 25.

Roy, A., & Vetterli, M. (2007). *The effective rank: A measure of effective dimensionality.* European Signal Processing Conference.

Templeton, A., et al. (2024). *Scaling and interpreting sparse autoencoders in language models.* Anthropic Technical Report.

Zheng, Z., et al. (2025). *RLVR is not RL: Geometric compression and alignment in reinforcement-trained language models.* Preprint.